Identifying influential multinomial observations by perturbation

Nyangoma, S.O.; Fung, W.-K.; Jansen, R.C.

Published in:
Computational Statistics %26 Data Analysis

DOI:
10.1016/j.csda.2005.04.023

Publication date:
2006

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 2799–2821
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Identifying influential multinomial observations by perturbation

S.O. Nyangoma[a,*], W.-K. Fung[b], R.C. Jansen[a]

[a] *Groningen Bioinformatics Centre, University of Groningen, Kerklaan 30, 9750 AA Haren, The Netherlands*
[b] *Department of Statistics & Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong*

Received 6 July 2004; received in revised form 29 April 2005; accepted 30 April 2005
Available online 31 May 2005

Abstract

The assessment of the influence of individual observations on the outcome of the analysis by perturbation has received a lot of attention for situations in which the observations are independent and identically distributed. However, no methods based on minor perturbations for carrying out such assessments are available in the context of multinomial models. A simultaneous perturbation scheme for the cell probabilities is proposed that leads to the definition of some new diagnostic tools for identifying influential observations. It is shown that the diagnostics derived extend and complement those based on the case deletion approach. The new diagnostics are used to explain departures from certain multinomial log-linear model assumptions. These tools are also used to give insights into genetic data for paternity.

Keywords: Perturbation; Diagnostics; Influence; Maximum likelihood estimate; Likelihood displacement; Conditional model

1. Introduction

The problem of identifying influential observations has received attention in recent years. The identification is usually done by introducing some perturbation in the problem formulation, and monitoring how these perturbations change the outcome of the analysis, such as parameter estimates, fitted values and goodness-of-fit measures. In linear and logistic regres-

* Corresponding author. Tel.: +31 50 363 39 57; fax: +31 50 363 38 00.
E-mail address: s.o.nyangoma@rug.nl (S.O. Nyangoma).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.04.023

sions, two approaches that have been used to quantify the effect of individual observations on these aspects of the fit are:

1. assessment by deletion (Andrews and Pregibon , 1978; Belsley et al., 1980; Cook, 1977, 1979; Cook and Weisberg, 1979; Pregibon, 1981),
2. assessment by infinitesimal perturbations (Belsley et al., 1980; Pregibon, 1981).

These approaches involve perturbing some metrics such as log-likelihood by allowing different weights to its components. Case deletion is a special case of the perturbation approach where all the cases are given weight 1, but the case of interest is given 0 weight. These approaches are directly applicable to other exponential family models in which the observations are independent and identically distributed (Tang et al., 2002; Lee and Xu, 2004). Because of the independence of the components of the metrics from which the diagnostics are constructed, the impact of individual observations can be precisely quantified by merely removing a term from a metric corresponding to the case of interest. However, in multinomial models, the terms in the log-likelihood function corresponding to the cells are not independent. Thus it does not make sense to merely perturb a term in the likelihood function. Suitable perturbation schemes are the simultaneous perturbations of the cell probabilities that take into account their dependence and at the same time and lead to perturbation of the cell observation whose influence is of interest in a unique manner in the estimation process. For example, in the case of the conditional multinomial approach, all the cell probabilities of the usual (unperturbed) multinomial model have been simultaneously modified in a manner that leads to the omission of the contribution of ith cell observation in the likelihood function (see Theorem 1). This approach has successfully been used to construct case "deletion" influence diagnostics analogous to those of the models in which observations are independent. However, no methods based on minor perturbations for quantifying the influence of individual cells on the fit of a multinomial model are available. When perturbing a model, a key requirement is that the parameters of the perturbed model should retain their original meanings and importance, and this is achieved by the proposed scheme.

We propose a simultaneous perturbation scheme for the cell probabilities of a multinomial model that leads to simple perturbation of only the ith cell (the cell whose influence we want to investigate) of the unperturbed multinomial model, in a manner that can be used to evaluate the impact of that cell on the analysis. The advantage of the proposed scheme is that it conserves the dependence of the cell probabilities. The perturbation scheme is constructed by modifying the model of Basu and Basu (1998) who defined a contamination scheme to study particular multinomial models. We use the perturbation to construct new diagnostic tools for identifying influential multinomial observations such as influence curves, average influence function, and Cook's curvature. Further diagnostics based on the likelihood displacement, change in Pearson statistic and change in deviance are also derived. The abilities of these tools in identifying influential observations are evaluated. The importance of one-step estimates in constructing certain diagnostics is emphasized and its shortcomings pointed out. The diagnostic tools derived are shown to effectively identify the influential cases and thus complement the case deletion diagnostics. Similarities between the new method and the conditional approach (Andersen, 1992; Nyangoma, 2000)

are stressed and the more specific nature of the new method pointed out. The new diagnostics are used to give insights into the categorical data that are modelled as log-linear and genetic data for paternity. It turns out that these diagnostics are functions of the basic building blocks of the influence diagnostics, as is the case with analogous diagnostics for models in which observations are independent. The critical use of the proposed scheme is the study infinitesimal departures from the usual multinomial model.

In the next section, we review the influence diagnostics for the multinomial models based on case deletion (Andersen, 1992; Nyangoma, 2000). The perturbation scheme is described in Section 3. We study the likelihood theory for the perturbed model in Section 4. We then derive the diagnostics for influential cells in Section 5. We explain how the diagnostics are derived for log-linear models in Section 6 and give applications of these diagnostics in Section 7. Finally, a discussion of the results is given in Section 8.

2. The multinomial model and the case deletion diagnostics

We assume that we have m-dimensional categorical data, $\boldsymbol{y}$, that are sufficiently modelled by the parametric multinomial model

$$L(\boldsymbol{y}, \boldsymbol{\pi}) = \frac{n!}{\prod_{j=1}^{m} y_j!} \prod_{j=1}^{m} \pi_j^{y_j}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ is an $m \times 1$ vector of the cell responses $y_i, i = 1, \ldots, m$ with the corresponding vector of probabilities $\boldsymbol{\pi} \in \mathbb{R}^m$, whose elements π_i are seen as functions $\pi_i(\boldsymbol{\theta})$ of a $p \times 1$ unknown parameter vector $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ and $p \leqslant m - 1$, $\mathbf{1}_m^{\mathrm{T}} \boldsymbol{y} = n$, $\mathbf{1}_m^{\mathrm{T}} \boldsymbol{\pi} = 1$, $\mathbf{1}_m$ denotes an $m \times 1$ vector of 1's and the superscript T denotes vector transpose. It is sufficient to assume that Θ is an open subset of $\mathbb{R}^p$.

To determine the influence of the ith observation on the various aspects of the fit of a multinomial model, Andersen (1992) proposed the use of the conditional multinomial model

$$L_i(\boldsymbol{y}, \boldsymbol{\pi}) = \frac{(n - y_i)!}{\prod_{j \neq i}^{m} y_j!} \prod_{j \neq i} \left(\pi_{j(i)}\right)^{y_j}, \tag{2}$$

in which the ith observation has been "deleted" and the cell probabilities for the remaining cells replaced with $\pi_{j(i)} = \pi_j / (1 - \pi_i)$, the conditional probability of observing an individual in cell j given that it is not observed in cell i.

Theorem 1. *The maximum likelihood (ML) estimates of the conditional multinomial model are equivalent to the ML estimates of an unperturbed multinomial model in which the original model assumptions concerning the probability of belonging to rth cell has been modified to*

$$\pi_r^c = \alpha_r \pi_r, \tag{3}$$

where the weights α_r are defined by

$$\alpha_r = \begin{cases} \dfrac{1}{1-\pi_i} & \text{if } r \neq i, \\ \dfrac{1}{\pi_i} & \text{if } r = i. \end{cases} \tag{4}$$

This perturbation gives rise to a likelihood function in which the contribution of the ith cell has been annihilated.

Proof. Note that the part of the conditional likelihood that is relevant in the estimation of parameters (kernel of the likelihood) may be expressed as

$$L_{ik} = \prod_{j \neq i} \left(\frac{\pi_j}{(1-\pi_i)} \right)^{y_j} = \left\{ \prod_{j \neq i} \left(\frac{\pi_j}{(1-\pi_i)} \right)^{y_j} \right\} \left(\frac{\pi_i}{\pi_i} \right)^{y_i} = \prod_{r=1}^{m} \left(\pi_r^c \right)^{y_r}, \tag{5}$$

where π_r^c is defined by 3 and 4. Thus the parameter estimates in the conditional model are equivalent to the parameter estimates obtained from a multinomial model in which the assumptions about the probabilities of belonging to given cells have been modified by giving similar weights to all cell probabilities but the probability of belonging to the *i*th cell is given a different weight that annihilates its contribution to the likelihood. □

Theorem 1 indicates that model perturbation, if chosen carefully can lead to data perturbation in a manner that can be used to evaluate the influence of individual observations on the fit of a model. This means that model perturbation offers a viable alternative to the usual practice of data perturbation.

It may be important to point out that the perturbation of the assumptions concerning the distribution of the error corresponding to particular observations (model perturbation) has successfully been used in regression to investigate the influence of those observations on the analysis. This is usually achieved by assigning equal weights (often ones) to all the error variances, but the error variance of the case whose influence is to be investigated is usually given a weight that is a reciprocal of some number that lies between zero and one. Thus the perturbation due to conditioning on particular cells follows the usual protocol as those in regression.

The perturbation due to conditioning on particular cells leads to a weighted parameter estimation procedure in which only the contribution of the cell of interest is penalized. This means that the difference between the ML estimates for the conditional model and the unconditional models provide information about the influence of the *i*th cell on the estimation. Let $\boldsymbol{D}$ be an $m \times m$ diagonal matrix with the *i*th diagonal element π_i. Define the vector $\boldsymbol{e}$ to have *i*th element e_i. Let $\boldsymbol{F}$ be an $m \times p$ derivative matrix of rank p with elements $f_i^r = \partial \pi_i / \partial \theta_r$ $(i = 1, \ldots, m, r = 1, \ldots, p)$ and $\boldsymbol{W}_{(i)}$ be a diagonal matrix whose diagonal elements are $w_{jj} = 1$ if $j \neq i$ and w_{ii}. Now, suppose $\{\boldsymbol{X}\}(\boldsymbol{\theta})$ denotes an $\boldsymbol{X}$ evaluated at $\boldsymbol{\theta}$. It can be shown that $\tilde{\boldsymbol{\theta}}_{(i)}$, the ML estimate of $\boldsymbol{\theta}$ under the conditional model, may be obtained through the iteratively reweighted least squares (IRLS) procedure

$$\boldsymbol{\theta}_{(i)}^{(r+1)} = \left\{ \left(\boldsymbol{F}^{\mathrm{T}} \boldsymbol{D}^{-1/2} \boldsymbol{W}_{(i)} \boldsymbol{D}^{-1/2} \boldsymbol{F} \right)^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{D}^{-1/2} \boldsymbol{W}_{(i)} \boldsymbol{z} \right\} \left(\boldsymbol{\theta}_{(i)}^{(r)} \right), \tag{6}$$

where $z = \boldsymbol{D}^{-1/2}\boldsymbol{F}\boldsymbol{\theta} + \frac{1}{\sqrt{n}}\boldsymbol{D}^{-1/2}\boldsymbol{e}$ is a working variable (Andersen, 1992; Nyangoma, 2000). Using some results from Rao (1973, pp. 128–129), it is easy to show that, asymptotically, $z \sim \mathrm{N}\left(\boldsymbol{D}^{-1/2}\boldsymbol{F}\boldsymbol{\theta},\ \left\{\boldsymbol{I}_m - \sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^{\mathrm{T}}\right\}\Big/ n\right)$.

The parameter improvement $\boldsymbol{\theta}_{(i)}^{(r+1)}$ is just a vector of coefficients from a weighted regression of $\{z\}\left(\boldsymbol{\theta}_{(i)}^{(r)}\right)$ on the columns of $\left\{\boldsymbol{D}^{-1/2}\boldsymbol{F}\right\}\left(\boldsymbol{\theta}_{(i)}^{(r)}\right)$ using weights from $\left\{\boldsymbol{W}_{(i)}\right\}\left(\boldsymbol{\theta}_{(i)}^{(r)}\right)$. The ML estimate is $\tilde{\boldsymbol{\theta}}_{(i)} = \lim_{r\to\infty}\boldsymbol{\theta}_{(i)}^{(r)}$. Nyangoma (2000) studies the properties of $\tilde{\boldsymbol{\theta}}_{(i)}$. The ML estimate for the unperturbed model, $\hat{\boldsymbol{\theta}}$, (Seber and Nyangoma, 2000) can be obtained iteratively from (6) using $\boldsymbol{W}_{(i)} = \boldsymbol{I}$. By applying (6) once to $\hat{\boldsymbol{\theta}}$ we obtain an expression for the one-step estimates (Andersen, 1992; Nyangoma, 2000). How the perturbation induced by conditioning affects the estimation is expressed by the structure of $\boldsymbol{W}_{(i)}$, in which the only weight that is different from one in the IRLS process is that corresponding to the ith cell. This means that the magnitude of the influence of ith cell on the analysis can be obtained by comparing $\tilde{\boldsymbol{\theta}}_{(i)}$ with $\hat{\boldsymbol{\theta}}$ (Andersen, 1992; Nyangoma, 2000). Some important metrics that have been used to measure the differences (distances) between $\tilde{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$ include the changes in the Pearson residuals, the Pearson goodness-of-fit statistic (Pearson distance), the deviance and the likelihood (the so-called likelihood distance) (Andersen, 1992; Nyangoma, 2000). Cases resulting in large distances are considered influential. All these diagnostics share the following properties: (1) directly proportional to the standardized residuals, $\hat{r}_{Pi} = \hat{e}_i/\sqrt{\hat{\pi}_i}$, (2) inversely proportional to the estimated variance of $\hat{r}_{Pi}$, $\hat{v}_i = \mathrm{var}\left(\hat{r}_{Pi}\right)$, and (3) monotonic increasing functions of the leverage, $\hat{h}_{ii}$, which is the ith diagonal element of $\boldsymbol{H} = \boldsymbol{D}^{-1/2}\boldsymbol{F}\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{F}\right)^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1/2}$, the orthogonal projection onto $\mathscr{C}\left(\boldsymbol{D}^{-1/2}\boldsymbol{F}\right)$, the subspace spanned by columns of $\boldsymbol{D}^{-1/2}\boldsymbol{F}$. Thus $\hat{r}_{Pi}, \hat{v}_i \approx (1-\hat{\pi}_i-\hat{h}_{ii})$ and $\hat{h}_{ii}$ are the basic building blocks for the influence diagnostics for the multinomial models. It is shown that the diagnostics derived from the method proposed in this manuscript are functions of these basic building blocks, a property that is characteristic of the influence diagnostics for models in which the observations are independent (see e.g. Pregibon, 1981).

3. The perturbation scheme

In the last section, it was established that the conditional multinomial model is equivalent to an unperturbed model in which the probabilities belonging to a cell has been perturbed in a manner that annihilates the contribution of the ith cell observation to the likelihood function using the scheme defined by 3 and 4. In the IRLS estimation process, that perturbation assigns a fixed weight to the cell under investigation, while the rest of the cells are not weighted. This implies that it cannot be used to study the effect of small modifications to the cell observations. This paper proposes a scheme that can be used to study the influence of a multinomial observation on the analysis by applying small changes to the assumptions about the probability of belonging to that cell. Suppose that the cell probabilities of the unperturbed model (1) are perturbed so that they are instead specified by

$$\pi_j^* = w_j\pi_j, \tag{7}$$

where w_j is defined by

$$w_j = \begin{cases} \lambda & \text{if } j \neq i, \\ \lambda + w/\pi_i & \text{if } j = i, \end{cases} \tag{8}$$

where $\lambda = (1 - w)$, $0 \leqslant w < 1$, $i, j = 1, \ldots, m$. Then $\sum_j \pi_j^* = 1$, so that the model takes into account the dependence of cell counts. This scheme defines a simultaneous perturbation of the cell probabilities in which $w_j = h(w)$, a function of w. In the next section, it is shown that this scheme has many attractive features that makes it suitable for studying the influence of individual cells on estimation. For example, as is the case with the conditional approach, it is shown that it gives rise to an IRLS estimation process in which the weights corresponding to all the cells are ones but that corresponding to the ith cell is $\phi_i(w) = \lambda\pi_i / (\lambda\pi_i + w)$, a property that motivates the use of the differences between the ML estimates for the perturbed and unperturbed models in quantifying the influence of the cell of influence on the estimation. This paper focuses on the induced perturbation $\phi_i(w)$ and not on w itself. Note that $w = 0$ corresponds to the null perturbation and gives rise to the unperturbed model (1). As pointed out by a referee, one may define perturbation schemes other than the one studied in this paper. Theorem 2 suggests an equivalent perturbation, which is described by Eqs. (11) and (12). Another alternative perturbation is explored in the discussion section of this paper.

4. The likelihood theory

On finding the ML estimates for the perturbed multinomial model proposed in the last section, we found it convenient to use the following theorem.

Theorem 2. *For every w, the ML estimates of the perturbed multinomial model whose cell probabilities are defined by Eqs.* (7) *and* (8) *are equivalent to the ML estimates of an unperturbed multinomial model in which the original model assumptions concerning the probability of belonging to ith cell has been modified to $\pi_i^* = (1 - w)\pi_i + w$. This model is equivalent to a multinomial model in which the ith observation is perturbed by a weight defined by $\vartheta_i(w) = log\left(\pi_i^*\right) / log\left(\pi_i\right)$.*

Proof. Let $L(\boldsymbol{\theta}|w)$ be the likelihood function (1) with the cell probabilities replaced with perturbed probabilities (7). Its log-likelihood is given by

$$\ell(\boldsymbol{\theta}|w) = \log\left(\frac{n!}{\prod_j y_j!}\right) + (n - y_i)\log(1 - w) + \sum_{j \neq i} y_j \log \pi_j + y_i \log \pi_i^*, \tag{9}$$

and for every w, the kernel of this log-likelihood is

$$\begin{aligned} \ell_{ik} &= \sum_{j \neq i} y_j \log \pi_j + y_i \log \pi_i^* \\ &= \sum_{j \neq i} y_j \log \pi_j + \vartheta_i(w) y_i \log \pi_i, \end{aligned} \tag{10}$$

which may be interpreted as the kernel of the usual multinomial model in which the assumptions concerning the probability of the ith cell membership has been modified to π_i^* or equivalently, it may be interpreted as the usual multinomial model in which the ith cell observation has been perturbed to $y_i^* = \vartheta_i(w)y_i$. □

Thus unlike the conditional model that deletes the ith cell observation, our scheme penalizes the contribution of this observation by a factor that is a function of w, which is defined by $\vartheta_i(w)$, with $w = 0$ representing the null perturbation. This weighting procedure is quite similar to those of regression, meaning that this scheme may be used to study the influence of the ith cell observation on the analysis. Note that the kernel described in Eq. (10) also represents a perturbation of the original model with the cell probabilities defined by

$$\pi_j^+ = w_j\pi_j, \tag{11}$$

where w_j is now defined by

$$w_j = \begin{cases} 1 & \text{if } j \neq i, \\ \lambda + w/\pi_i & \text{if } j = i. \end{cases} \tag{12}$$

This perturbation is consistent with the perturbation of error variance in regression where only the assumption regarding the distribution of the ith error is modified. So the question that now arises is: Can this perturbation be used to construct diagnostic tools for influential multinomial observations? The results in the rest of this manuscript attempt to answer this question in affirmative. It may be established that Eqs. (3), (4), (7), (8), (11) and (12) describe a family of perturbations for multinomial models that involve modifying the model assumptions about the probability of cell by assigning them weights that are all functions $h(w)$, of w, with the perturbation due to conditioning on certain cells being the identity case of $h(w)$. Thus, this paper unifies the perturbation approaches that are used to study the influence individual cell observations.

Since $\ell(\boldsymbol{\theta}|w)$ is at least twice differentiable with respect to $(\boldsymbol{\theta}, w)$ and $\ell(\boldsymbol{\theta}|w=0) = \ell(\boldsymbol{\theta})$, the log-likelihood for the unperturbed model, our scheme possesses the properties of the general perturbation schemes such as those proposed by Cook (1986) and Escober and Meeker (1992).

As in the case of the conditional model and in the regression models, a justification for the use the proposed perturbation in assessing the influence of individual multinomial observations, may be deduced from the ML estimation process. It can be shown that the score function for the perturbed model $\boldsymbol{u}_w$ is given by

$$\boldsymbol{u}_w = \sqrt{n}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1/2}\boldsymbol{W}_{(i)}^w\boldsymbol{D}^{-1/2}\boldsymbol{e}_w, \tag{13}$$

where $\boldsymbol{e}_w$ is an $m \times 1$ vector with elements $e_{wj} = \left(y_j - n\pi_j^*\right)/\sqrt{n}$ and $\boldsymbol{W}_{(i)}^w$ is a diagonal matrix with $\phi_i(w)$ in the ith diagonal position and 1's elsewhere. The rth element of $\boldsymbol{u}_w$ is

$$u_{wr} = \sum_{j\neq i}\left(y_j - n\pi_j^*\right)\frac{1}{\pi_j}\frac{\partial\pi_j}{\partial\theta_r} + \phi_i(w)\left\{y_i - n\pi_i^*\right\}\frac{1}{\pi_i}\frac{\partial\pi_i}{\partial\theta_r}. \tag{14}$$

For a given value of π_i, $\phi_i(w)$ is a monotonic decreasing function of w. Then the contribution of the ith cell to the magnitude of the components of $\boldsymbol{u}_w$ decreases as w approaches 1, denoted by $w \to 1$.

The expected Fisher information for the perturbed model is given by

$$\boldsymbol{I}_w = n\lambda \boldsymbol{F}^{\mathrm{T}} \boldsymbol{D}^{-1/2} \boldsymbol{W}_{(i)}^{w} \boldsymbol{D}^{-1/2} \boldsymbol{F}, \tag{15}$$

and its (r, s)th element is

$$I_w^{sr} = E\left(-\frac{\partial^2 \ell_w}{\partial\theta_s \partial\theta_r}\right) = \sum_{j\neq i} n\lambda \frac{\partial\pi_j}{\partial\theta_s} \frac{1}{\pi_j} \frac{\partial\pi_j}{\partial\theta_r} + n\lambda\phi_i \frac{\partial\pi_i}{\partial\theta_s} \frac{1}{\pi_i} \frac{\partial\pi_i}{\partial\theta_r}. \tag{16}$$

Note that as $w \to 1$, $\lambda \to 0$ and $\phi_i(w) \to 0$, so that the last term in (16) approaches 0 faster than the first, indicating that the contribution of the ith cell on the magnitude of the elements of $\boldsymbol{I}_w$ also reduces as $w \to 1$.

For a fixed w, the usual regularity conditions (Birch, 1964) hold and the ML estimates for the perturbed model can be obtained uniquely by the method of IRLS, which is defined by

$$\boldsymbol{\theta}_w^{(r+1)} = \left\{\left(\boldsymbol{F}^{\mathrm{T}} \boldsymbol{D}^{-1/2} \boldsymbol{W}_{(i)}^{w} \boldsymbol{D}^{-1/2} \boldsymbol{F}\right)^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{D}^{-1/2} \boldsymbol{W}_{(i)}^{w} \boldsymbol{z}_w\right\} \left(\boldsymbol{\theta}_w^{(r)}\right), \tag{17}$$

where

$$\boldsymbol{z}_w = \boldsymbol{D}^{-1/2} \boldsymbol{F}\boldsymbol{\theta} + \frac{1}{\sqrt{n}\lambda} \boldsymbol{D}^{-1/2} \boldsymbol{e} + \frac{w}{\lambda} \boldsymbol{D}^{-1/2} \left(\boldsymbol{\pi} + \boldsymbol{q}_{(i)}\right)$$

is a working variable for the perturbed model, $\boldsymbol{q}_{(i)}$ is an $m \times 1$ vector with elements $q_{(i)j} = 0$, $j \neq i$ and $q_{(i)i} = -1$. The parameter improvement $\boldsymbol{\theta}_i^{(r+1)}(w)$ is then a weighted regression of $\{\boldsymbol{z}_w\}\left(\boldsymbol{\theta}_i^{(r)}(w)\right)$ on the columns of $\left\{\boldsymbol{D}^{-1/2}\boldsymbol{F}\right\}\left(\boldsymbol{\theta}_i^{(r)}(w)\right)$ with the diagonal elements of $\left\{\boldsymbol{W}_{(i)}^{w}\right\}\left(\boldsymbol{\theta}_i^{(r)}(w)\right)$ as weights. The ML estimate for the perturbed model is $\tilde{\boldsymbol{\theta}}_i(w) = \lim_{r\to\infty} \boldsymbol{\theta}_i^{(r)}(w)$.

How our perturbation affects the estimation process may also be deduced from the structure of $\boldsymbol{W}_i^w$, in which the only weight that is different from one is that corresponding to the ith cell. This indicates that the differences between the ML estimates for the perturbed and the unperturbed models may also provide information regarding the impact of the ith cell on the fit of a multinomial model. For $w = 0$, $\phi_i(0) = 1$ and $\lambda = 1$ and (13) and (15) give the corresponding expressions for the unperturbed model and (17) is an iterative method for obtaining $\hat{\boldsymbol{\theta}}$. Then the influence of the ith cell observation may be assessed by comparing $\hat{\boldsymbol{\theta}}$ with $\tilde{\boldsymbol{\theta}}_i(w)$.

To be able to construct multinomial analogues of linear regression diagnostics, we make use of the one-step ML estimates

$$\tilde{\boldsymbol{\theta}}_i^1(w) = \hat{\boldsymbol{\theta}} - \frac{w\left(\hat{\boldsymbol{F}}^{\mathrm{T}} \hat{\boldsymbol{D}}^{-1} \hat{\boldsymbol{F}}\right)^{-1} \hat{\boldsymbol{g}}_i \hat{r}_{Pi}}{\sqrt{n}\lambda\left(\hat{\pi}_i + w\hat{v}_i\right)} - \frac{w\hat{\pi}_i^{1/2}\left(\hat{\boldsymbol{F}}^{\mathrm{T}} \hat{\boldsymbol{D}}^{-1} \hat{\boldsymbol{F}}\right)^{-1} \hat{\boldsymbol{g}}_i}{\lambda\left(\hat{\pi}_i + w\hat{v}_i\right)}, \tag{18}$$

where $\hat{r}_{Pi}$ is the ith standardized residual (Haberman, 1973) $\hat{v}_i = 1 - \hat{\pi}_i - \hat{h}_{ii}$, $\boldsymbol{g}_i^{\mathrm{T}}$ is the ith row of $\boldsymbol{D}^{-1/2}\boldsymbol{F}$, and the hats indicate evaluation at $\hat{\boldsymbol{\theta}}$. The estimate $\tilde{\boldsymbol{\theta}}_i^1(w)$ is obtained by taking the initial value in the iterative scheme (17) as $\tilde{\boldsymbol{\theta}}_i(0) = \hat{\boldsymbol{\theta}}$. The standardized residuals are used to diagnose extreme cells, while the diagonal elements of $\boldsymbol{H}$ have been used (Andersen, 1992; Seber and Nyangoma, 2000) as measures of leverages in multinomial models. It is easy to see that $\tilde{\boldsymbol{\theta}}_i^1(0) = \hat{\boldsymbol{\theta}}$. As $\lim_{w\to 1}(\hat{\pi} + w\hat{v}_i) = \left(1 - \hat{h}_{ii}\right)$, we find that the last two terms on the right-hand side of (18) will be quite large if the perturbed observation is influential as $w \to 1$. This means that $\tilde{\boldsymbol{\theta}}_i^1(w)$ can be partitioned as a sum of the ML estimates for the unperturbed model plus correction term containing diagnostic information about the influence of the ith cell (i.e. a function of the basic building blocks for the influential multinomial observations) and thus have similar characteristics to the one-step estimates for other models such as logistic regression (Pregibon, 1981), linear regression (Cook and Weisberg, 1982) and survival analysis (Cain and Lange, 1984). The one-step estimates in those models have been used to define influence curves. Here, we make use of $\tilde{\boldsymbol{\theta}}_i^1(w)$ to derive several diagnostics for influential multinomial observations. We see that the influence of the ith cell on the model may be defined as $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_i^1(w)$. The accuracy of the influence curve can be evaluated by comparing it with the exact difference $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_i(w)$. It is often useful to take $w = \bar{w}$, where $0 < \bar{w} < 1$.

Suppose $h\left(\hat{\boldsymbol{\theta}}\right)$ is a function of $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$, and since $\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = \mathrm{O}_P(1)$, then to the order of linear approximation, $h\left(\hat{\boldsymbol{\theta}}\right) = h(\boldsymbol{\theta}) + \mathrm{O}_P(1/\sqrt{n})$, where P denotes probability and $h\left(\hat{\boldsymbol{\theta}}\right) \approx h(\boldsymbol{\theta})$. Now, since $\hat{r}_{Pi} \approx N(0, 1 - \pi_i - h_{ii})$ and $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ almost surely, it follows from (18) that $\tilde{\boldsymbol{\theta}}_i^1(w)$ is an asymptotically biased estimate for $\boldsymbol{\theta}$ when $w > 0$. Its bias is

$$b(\boldsymbol{\theta}) = E\left[\tilde{\boldsymbol{\theta}}_i^1(w)\right] - \boldsymbol{\theta} = -\frac{w\sqrt{\pi_i}\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{F}\right)^{-1}\boldsymbol{g}_i}{\lambda(\pi_i + wv_i)}, \tag{19}$$

which can be large if the ith cell is influential and it is close to zero for small w. This explains why $\tilde{\boldsymbol{\theta}}_i^1(w)$ tends to exaggerate the influence as we will see in the numerical examples in Section 7. We must point out that this property does not however, undermine the use of the one-step estimates as tools for constructing influence diagnostics as demonstrated in the numerical examples in Section 7. However, it may be better to use the adjusted one-step estimates $\tilde{\boldsymbol{\theta}}_{ai}^1(w) = \tilde{\boldsymbol{\theta}}_i^1(w) - b\left(\hat{\boldsymbol{\theta}}\right)$, that are asymptotically unbiased. Note that the one-step estimates are asymptotically unbiased when $w \approx 0$. Because $\hat{\boldsymbol{\theta}}$ and $\hat{r}_{Pi}$ are asymptotically independent, one can show that the variance of $\tilde{\boldsymbol{\theta}}_i^1(w)$ exceeds the Cramer–Rao lower bound by an amount

$$b_v(\boldsymbol{\theta}) = \frac{w^2 v_i\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{F}\right)^{-1}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{F}\right)^{-1}}{n\lambda^2(\pi_i + wv_i)^2}, \tag{20}$$

which is 0 for large n or when $w \approx 0$, but can be large in presence of influential observations, for $w > 0$. This indicates that if the perturbed model is the true model but one uses the usual multinomial model (unperturbed), the conclusions can be quite erroneous.

5. Diagnostics for influential observations

In this section, we derive diagnostic tools that may be used to quantify the influence of single cells on the analysis.

5.1. The influence curves

We begin by constructing measures of the local influence, which involves monitoring the changes in the parameter estimates when cell probabilities are slightly perturbed through the scheme defined by (7) and (8). Following Cook (1986) and Shi (1997), we can define the influence function for $\hat{\boldsymbol{\theta}}$ as

$$\frac{\partial \tilde{\boldsymbol{\theta}}_i^1(w)}{\partial w} = \Delta_i(w) = \frac{\left(\hat{\pi}_i + w^2\hat{v}_i\right)\left(\hat{\boldsymbol{F}}^{\mathrm{T}}\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{F}}\right)^{-1}\hat{\boldsymbol{g}}_i\hat{r}_{Pi}}{\sqrt{n}\lambda^2\left(\hat{\pi}_i + w\hat{v}_i\right)^2} + \frac{\left(\hat{\pi}_i + w^2\hat{v}_i\right)\left(\hat{\boldsymbol{F}}^{\mathrm{T}}\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{F}}\right)^{-1}\hat{\boldsymbol{g}}_i\hat{\pi}_i^{1/2}}{\lambda^2\left(\hat{\pi}_i + w\hat{v}_i\right)^2}. \quad (21)$$

Since $\lim_{w\to 1}\left(\hat{\pi}_i + w^2\hat{v}_i\right)/\left(\hat{\pi}_i + w\hat{v}_i\right)^2 = \left(1 - \hat{h}_{ii}\right)^{-1}$ and the right-hand side of (21) is of $\mathrm{O}_P\left(\lambda^{-2}\right)$, $\Delta_i(w)$ will be quite large as $w \to 1$, especially if the perturbed observation is influential. Then $\Delta_i(0)$ may be seen as the influence on $\hat{\boldsymbol{\theta}}$ of perturbing cell probabilities through scheme (8) (Hampel et al., 1985; Shi, 1997). Useful diagnostic tools for identifying influential observations are the values of $\Delta_i(w)$ evaluated at some point $\bar{w}$, $0 < \bar{w} < 1$ (see e.g. Pregibon, 1981). Since $\tilde{\boldsymbol{\theta}}_i^1(w)$ is continuously differentiable in $0 \leqslant w < 1$, an application of the mean value theorem (MVT) yields

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_i^1(1) = -\Delta_i(\bar{w}). \quad (22)$$

This is the local change in coefficients when the ith cell is not given full weight.

Motivated by the definition of an influence measure based on sample influence curves for case deletion by (Cook and Weisberg, 1982, p. 110), we define a measure of influence as

$$CD = \{\Delta_i(0)\}^{\mathrm{T}}\boldsymbol{M}\{\Delta_i(0)\}/c, \quad (23)$$

where $\boldsymbol{M}$ is some $p \times p$ positive-definite matrix and c is a scalar. Shi (1997, p. 117) defined a related measure based on directional perturbation. Note that $\Delta_i(0)$ is a function of $\hat{r}_{Pi}$, and thus it contains similar information to the corresponding expression for the logistic regression (Pregibon, 1981).

Let $\alpha_i(w) = E[\Delta_i(w)]$ be the average influence function. Then to the order of linear approximation

$$\alpha_i(w) = \frac{(\pi_i + w^2 v_i)\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{F}\right)^{-1}\boldsymbol{g}_i\pi_i^{1/2}}{\lambda^2(\pi_i + w v_i)^2}, \tag{24}$$

which contains similar diagnostic information as $\Delta_i(w)$. Then $\alpha_i(w)$ may be considered as an influence function in its own right and can be used to construct metrics for influential observations.

5.2. Likelihood displacement

A more general way of comparing $\tilde{\boldsymbol{\theta}}_i(w)$ and $\hat{\boldsymbol{\theta}}$ is by their likelihood displacements defined by

$$LD_i(w) = 2\sum_j y_j\left(\log\hat{\pi}_j - \log\tilde{\pi}_j\right), \tag{25}$$

where the tildes at the top indicate an evaluation at $\tilde{\boldsymbol{\theta}}_i(w)$. The likelihood displacement has proved useful in identifying influential observations for a variety of models (Williams, 1987; Nyangoma, 2000; Fung and Kwan , 1997).

Then following Cook (1986), we define a curvature measure for multinomial models as

$$CC = \left.\frac{\partial^2 LD_i(w)}{\partial^2 w}\right|_{w=0} = -2\sum_{j=1}^{m} y_j\frac{\partial^2\log\left(\tilde{\pi}_j\right)}{\partial^2 w}, \tag{26}$$

where $\tilde{\pi}_j$ are the cell probabilities for the unperturbed model evaluated at $\tilde{\boldsymbol{\theta}}(w)$. Replacing $W_{(i)}$ and $\boldsymbol{U}^+$ with $\boldsymbol{I}$ and $\boldsymbol{U}$, respectively, in Eq. (4) of (Nyangoma, 2000), we find that the observed information for the unperturbed model is $\boldsymbol{M} = n\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{F} - \frac{1}{\sqrt{n}}\boldsymbol{B}\right)$, where $\boldsymbol{B} = \left[\boldsymbol{e}^{\mathrm{T}}\boldsymbol{D}^{-1}\right][\boldsymbol{G} - \boldsymbol{U}]$ is a $p \times p$ matrix, $\boldsymbol{U} = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{F}$, $\boldsymbol{V}$ is an $m \times m \times m$ array whose ith face is an $m \times m$ matrix with elements $v_i^{ii} = \pi_i$ and $v_i^{ij} = 0$, for $i \neq j$, $\boldsymbol{G}$ is an $m \times p \times p$ array with ith face being a $p \times p$ matrix with elements $f_i^{rs} = \partial^2\pi_i/\partial\theta_r\partial\theta_s$. It follows from Lawrance (1991) that a measure of Cook's curvature for multinomial models may be written as

$$CC = -\{\Delta_i(0)\}^{\mathrm{T}}\hat{\boldsymbol{M}}\{\Delta_i(0)\}, \tag{27}$$

which is basically the unscaled version of Cook's distance defined in (23), with $\boldsymbol{M}$ replaced with $\hat{\boldsymbol{M}}$. Note that, asymptotically, $\Delta_i(w) \propto \left(\tilde{\boldsymbol{\theta}}_i^1(w) - \hat{\boldsymbol{\theta}}\right)$, which justifies the use of $\Delta_i(w)$ as an influence curve (Cook and Weisberg, 1982, p. 182) and hence the use of CC and CD as metrics for the distance between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}_i^1(w)$. This suggests another metric of the form

$$LD_i(w) = \tilde{\boldsymbol{\delta}}_i(w)^{\mathrm{T}}\hat{\boldsymbol{M}}\tilde{\boldsymbol{\delta}}_i(w), \tag{28}$$

where $\tilde{\boldsymbol{\delta}}_i(w) = \left(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_i^1(w)\right)$, which is basically a second-order approximation to the likelihood displacement and may be seen as a version of Cook's original distance for perturbed

multinomial models. Note that $\boldsymbol{M}/n = \boldsymbol{I}_0 + \mathrm{O}_P\left(n^{1/2}\right)$, where $\mathrm{O}_P\left(n^{-1/2}\right)$ represents terms that are of order $n^{-1/2}$ and $\boldsymbol{I}_0 = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{F}$ is the scaled value of the expected Fisher information for the perturbed model evaluated at $w = 0$. In many applications, it may be enough to use the inner product $n\hat{\boldsymbol{I}}_0$ instead of $\hat{\boldsymbol{M}}$, in which case LD_i is a perturbation version of the unscaled Cook's distance for the case deletion reported in Andersen (1992). A graph of $LD_i(w)$ vs. w is the so-called influence graph. The magnitude of the curvature of such a graph can be used to diagnose the impact of the perturbation on the likelihood. The curvature information can be extracted using methods of local influence (Cook, 1986; Fung and Kwan , 1997).

5.3. Pearson distance

Let $\boldsymbol{e} = \boldsymbol{e}_0$, the vector of errors for perturbed model $\boldsymbol{e}_w$, evaluated at $w = 0$. Then $\boldsymbol{e}$ is a vector of errors for the unperturbed model and the statistic $\hat{S}$, where $S = \boldsymbol{e}'\boldsymbol{D}^{-1}\boldsymbol{e}$, is the Pearson goodness-of-fit statistic. Following Lee et al. (2002), $PD = S - \hat{S}$ and $\hat{S}$ have asymptotically independent χ^2 distributions with p and $m - p - 1$ degrees of freedoms, respectively, so that, asymptotically $FD = (m - p - 1)PD/p\hat{S} \sim F_{p,m-p-1}$. Suppose $\tilde{S} = S\left(\tilde{\boldsymbol{\theta}}_i(w)\right)$, then two important diagnostics for assessing the impact of perturbation (7) on the fit of a models are

$$PD_i(w) = \tilde{S} - \hat{S} \quad \text{and} \quad FD_i(w) = (m - p - 1)PD_i(w)/p\hat{S}. \tag{29}$$

The former will be called the Pearson distance, whereas the latter, the F-distance. It is easy to show that $PD_i(w) \approx \tilde{\boldsymbol{\delta}}_i^{\mathrm{T}}(w)\left(n\hat{\boldsymbol{I}}_0\right)\tilde{\boldsymbol{\delta}}_i(w)$, thus $PD_i(w)$ can be interpreted in much the same way as $LD_i(w)$. Then the one-step approximations of $PD_i(w)$ can be expressed as

$$PD_i^1(w) \approx \gamma_i(w)^2\left\{D_i^1 + h_{ii}^* q_i\right\}, \tag{30}$$

where $D_i^1 = \hat{h}_{ii}\hat{r}_{Pi}^2/pc\hat{v}_i$ is the Cook's distance for the multinomial models, $h_{ii}^* = \hat{h}_{ii}/\hat{v}_i$ is a monotonic increasing function of $\hat{h}_{ii}$ called the "potential" (Nyangoma, 2000), $\gamma_i(w) = w\sqrt{pc\hat{v}_i}/\lambda\left(\hat{\pi}_i + w\hat{v}_i\right)$, $c = \hat{S}/(m - p - 1)$, $q_i = a\hat{r}_{Pi} + b$, $a = 2\sqrt{n}\hat{\pi}_i/pc$ and $b = n\hat{\pi}_i/pc$. Asymptotically, this approximation is equal to the Cook's scalar measure of distance. It is easy to see that $LD_i(w) \approx PD_i(w)$. Since $\lim_{w\to 1}\left(\hat{\pi} + w\hat{v}_i\right) = \left(1 - \hat{h}_{ii}\right)$, $\gamma_i(w)$ can be appreciably large as $w \to 1$ if the perturbed cell is influential. Since $PD_i^1(w)$ is a function of $\gamma_i(w)$, P_i, $\hat{r}_{Pi}$, and $\hat{v}_i$ it will be affected by the influential observations in much the same way as these observations affect the case deletion diagnostics. Hence $PD_i^1(w)$ may be described globally by just two summary statistics, namely its curvature evaluated at $w = 0$ and D_i^1. A similar conclusion was deduced for linear regression models by Critchley (in a discussion of a paper by Cook, 1986). Fung (1992, 1995) also discussed the two summary statistics in a regression setting for discriminant analysis.

Following Weisberg (1985) an observation resulting in an $FD_i(w)$ that is close to 1 when perturbed is asymptotically considered influential.

5.4. Diagnostics based on the deviance.

The effect of case-deletion on the deviance for a multinomial model has been studied by (Nyangoma, 2000). Here, we develop diagnostics based on the deviance that may be used to evaluate the effect of infinitesimal perturbations on the fit of a multinomial model.

Let $p_h = y_h/n$, $h = 1, \ldots, m$, be the proportion of the subjects falling in the hth group. Then p_h maximizes the unconstrained unperturbed multinomial model. The deviance for the perturbed model is given by

$$D_w\left(\tilde{\boldsymbol{\theta}}_i(w)\right) = \tilde{d}_i + D\left(\tilde{\boldsymbol{\theta}}_i(w)\right), \tag{31}$$

where $\tilde{d}_i = \left(\tilde{r}_{D_w i}\right)^2 - (\tilde{r}_{Di})^2$, $\tilde{r}_{D_w i} = q_i^*\left\{y_i\left(\log p_i^* - \log \tilde{\pi}_i^*\right)\right\}^{1/2}$ is the ith deviance residual for the perturbed model, $\tilde{r}_{Dj} = q_j\left\{y_j\left(\log p_j - \log \tilde{\pi}_j\right)\right\}^{1/2}$ is the jth deviance residual for the unperturbed model with $\hat{\boldsymbol{\theta}}$ replaced by $\tilde{\boldsymbol{\theta}}_i(w)$, $q_j = sign\left(p_j - \tilde{\pi}_j\right)$, $q_i^* = sign\left(p_i^* - \tilde{\pi}_i^*\right)$, $p_i^* = \lambda p_i + w$, and $D\left(\tilde{\boldsymbol{\theta}}_i(w)\right) = \sum_j \tilde{r}_{Dj}$ is the deviance for the unperturbed model with $\hat{\boldsymbol{\theta}}$ replaced by $\tilde{\boldsymbol{\theta}}_i(w)$. This suggests that the change in the ith deviance residuals,

$$\tilde{d}_i = \left(\tilde{r}_{D_w i}\right)^2 - (\tilde{r}_{Di})^2, \tag{32}$$

is a possible measure of the impact of the ith cell on the analysis. The one-step approximation to this change may be obtained by replacing $\tilde{\boldsymbol{\theta}}_i^1(w)$ with $\tilde{\boldsymbol{\theta}}_i(w)$. The resulting statistic is similar to an influence diagnostic for the generalized linear models (GLM) (Williams, 1987). Useful versions of these diagnostics may be obtained by replacing w with $w = \bar{w}$, where $0 < \bar{w} < 1$.

An important use of the proposed perturbation scheme is the assessment of departures from the usual multinomial assumptions. These departures may be viewed as a tests of hypotheses problem in which the null hypothesis is $H_0 : w = 0$, that the unperturbed model adequately models the data, against the alternative $H_1 : w > 0$, that the perturbed model is the right model. This hypothesis is tested by comparing $D\left(\hat{\boldsymbol{\theta}}\right) - D_w\left(\tilde{\boldsymbol{\theta}}_i^1\left(\hat{w}\right)\right)$ with its null asymptotic χ^2_{p+1} distribution. The following Lemma gives a connection between this statistic and the likelihood distance.

Lemma. *The adjusted change in deviance*

$$\hat{x}_i(w) = D_w\left(\tilde{\boldsymbol{\theta}}_i^1(w)\right) - D\left(\hat{\boldsymbol{\theta}}\right) + \tilde{d}_i^1, \tag{33}$$

where $\tilde{d}_i^1$ is the one-step approximation to $\tilde{d}_i$ is asymptotically equivalent to the likelihood distance.

Proof. Find a Taylor series expansion of $D\left(\tilde{\boldsymbol{\theta}}_i(w)\right)$ about $\hat{\boldsymbol{\theta}}$ and substitute it in Eq. (31) to obtain

$$\hat{x}_i(w) = \frac{1}{2}\tilde{\boldsymbol{\delta}}_i^{\mathrm{T}}(w)\left(n\hat{\boldsymbol{I}}_0\right)\tilde{\boldsymbol{\delta}}_i(w). \qquad \square \tag{34}$$

The Lemma suggests that the corrected change in deviance, $\hat{x}_i(w)$, may be interpreted as the case perturbation analogue of the Cook's scalar measure of distance (Andersen, 1992). Index plot of this statistic may be used to identify influential cells. The $(1-\alpha)100\%$ confidence interval for $\boldsymbol{\theta}$ after perturbing the model as in (7) is approximated by the set of values w for which $\hat{x}_i(w) < \chi^2_1(\alpha)$. Further diagnostics are the differential values of $\hat{x}_i(w)$ as w departs from 0.

6. Perturbation diagnostics for multinomial log-linear models

The tools developed above may be adapted to study the effect of small perturbations of individual cell probabilities of the multinomial log-linear models, e.g. the logit model (Agresti, 1990 p. 313). For these models, the jth cell probability is

$$\pi_j = \frac{\exp\left(\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\theta}\right)}{\sum_k \exp\left(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\theta}\right)}, \tag{35}$$

where $\boldsymbol{x}_j^{\mathrm{T}}$ is the jth row of the model matrix, $\boldsymbol{X}$, for the multinomial log-linear models. For these models, the expressions for most of the components of the above diagnostics simplify (see e.g. Nyangoma, 2000). For example, the matrix $\boldsymbol{D}^{-1/2}\boldsymbol{F}$, that plays a key role in the construction of diagnostic tools reported in this paper, is given by

$$\boldsymbol{D}^{-1/2}\boldsymbol{F} = \left(\boldsymbol{I} - \sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^{\mathrm{T}}\right)\boldsymbol{D}^{1/2}\boldsymbol{X}. \tag{36}$$

The ML estimates of the parameters for perturbed model can be obtained through the IRLS (17) using $\boldsymbol{D}^{-1/2}\boldsymbol{F}$ defined by (36). For $w=0$, (17) is just the IRLS procedure for obtaining $\hat{\boldsymbol{\theta}}$, the ML estimate of a Poisson log-linear model. The matrix $\boldsymbol{X}$ is the design matrix of the same Poisson model with the first column omitted. Both $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{X}$ may be obtained by fitting the data using the glm object of R, with family specification, Poisson. The vector of Pearson residuals, that also features in most diagnostics presented here, are the residuals from this fit. Thus for log-linear models, many diagnostics are readily computed using the output from the fit of Poisson log-linear models.

7. Numerical examples

7.1. Genetic model for mother–child relationship

This model was investigated by Hirschfeld and Heiken (1963) (see Nyangoma, 2000; Seber and Nyangoma, 2000; Elandt-Johnson, 1971 p. 326) and concerns the codominant allele blood group system Gc. This system has two alleles Gc^1 and Gc^2, with phenotypes denoted by $Gc1-1$, $Gc2-1$ and $Gc2-2$. If the gene probabilities for Gc^1 and Gc^2 are, respectively, θ and $1-\theta$, then the $3\times 3 = 9$ mother–child genotype combinations have a multinomial distribution with joint probabilities reported in Elandt-Johnson (1971, p. 325, Table 12.3). The MLE of θ based on the Hirschfeld and Heiken (1963) data, consisting of

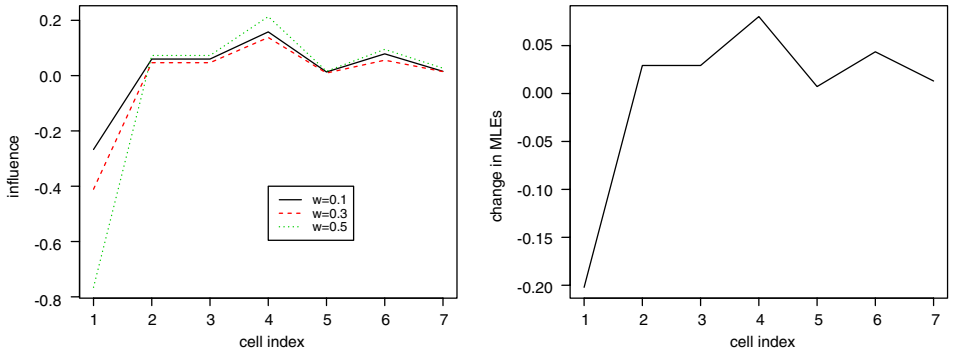

Fig. 1. Influence curves (left) and change in MLEs at $w = 0.5$ (right).

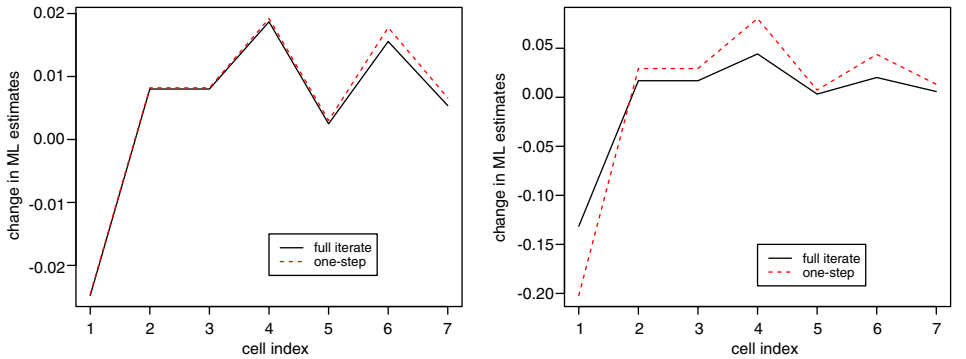

Fig. 2. Change in MLEs at $w = 0.1$ (left) and $w = 0.5$ (right).

$n = 142$ mother–child pairs, is $\hat{\theta} = 0.231$. The case deletion diagnostics applied to this data set by Nyangoma (2000) identified cell 1 as influential.

We applied some diagnostics proposed in this paper to the Hirschfeld and Heiken (1963) data. Fig. 1 displays the influence curves for $\bar{w} = 0.1$, 0.3 and 0.5 and the changes in ML estimates. It is clear from this figure that a perturbation of Cell 1 results in the largest values of $\Delta_i(\bar{w})$ and $\hat{\theta} - \tilde{\theta}_i^1(w)$, while a perturbation of the other cells produce small changes. These changes decrease in magnitude but remain rather constant for the other cells as $w \to 0$. It is then sensible to conclude that Cell 1has the most influence of all cells.

Fig. 2 displays the changes in the ML estimates for $\bar{w} = 0.1$ and 0.5 based on the one-step and full iterated estimates. For w close to 0, the changes in ML estimates are quite similar for both the one-step and the full iterated estimates. However, these estimates can be very different for large w, with one-step estimates often exaggerating the displacement as demonstrated by the plot at $w = 0.5$. Although these differences exist, likelihood displacements based on both the estimates (Fig. 3) do indicate that cell 1 has the largest effect when perturbed. It is clear that large values of w lead to better display of the influence. Evaluation

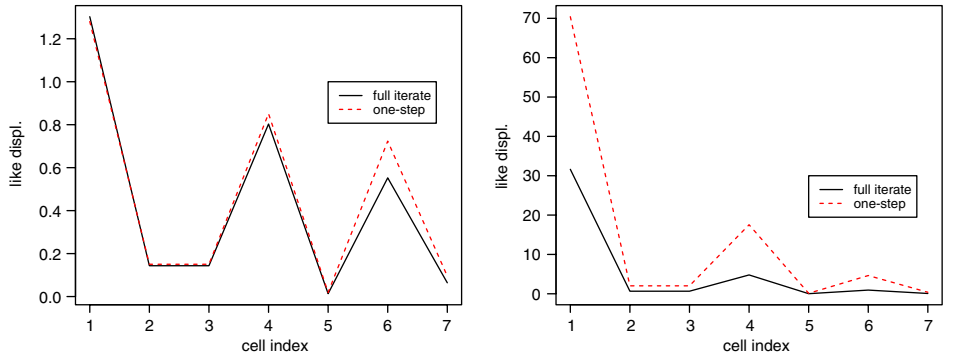

Fig. 3. Likelihood displacements (Eq. (25))—$w = 0.1$ (left) and $w = 0.5$ (right).

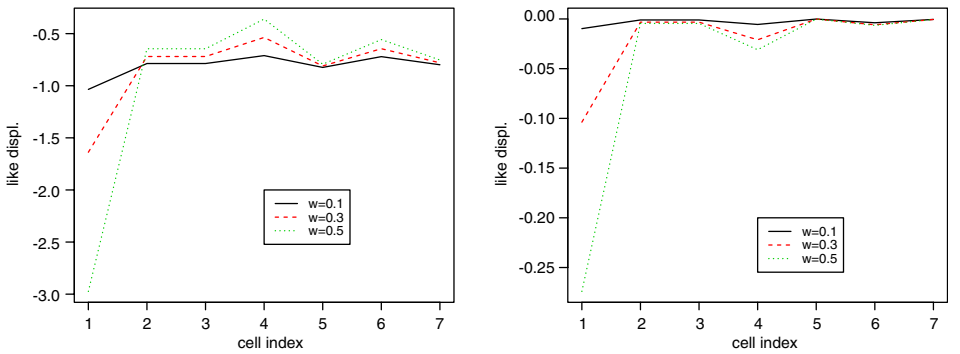

Fig. 4. Likelihood displacements approx (Eq. (28))—one-step MLE (left) and full iteration MLE (right).

of the other approximations to the likelihood displacement (Eq. (25)) at both the one-step and full iterated estimates at various values of w also identifies cell 1 as the most influential (see Fig. 4). The magnitudes of these changes increase as $w \to 1$. Cook's curvature (Fig. 5) gave similar conclusions for $w = 0.1, 0.3, 0.5$. Figs. 6 and 7 display the so-called Pearson distance computed at both the full iterated and the one-step estimates. For $w = 0.1$, the Pearson displacement does not capture cell 1 as the most influential. However, for large w ($w > 0$), this statistic gives similar conclusions as the likelihood displacement. The changes in the deviance (Eq. (32)) gave similar conclusions as those based on the Pearson distance as can be seen in Figs. 8 and 9. It is noteworthy that the modified change in deviance (Figs. 10 and 11) also effectively identifies the influential cell.

In conclusion, for large w ($w > 0$), there seems to be an agreement between the one-step and the full iterative diagnostics that Cell 1 is the most influential among all the cells. The ability to find influential cells can benefit the analyst in at least two ways. First, the study of influence yields information concerning the reliability of the conclusions and their dependence on the assumed model. Second, the outlying cells will tend to have, on average,

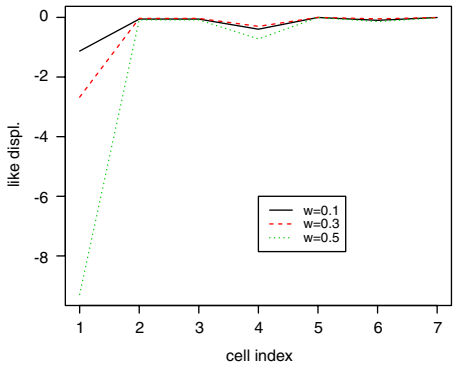

Fig. 5. Cook's curvature (Eq. (27)).

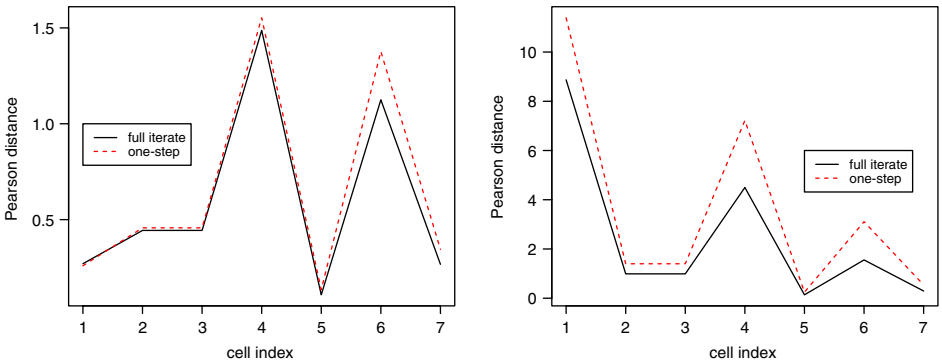

Fig. 6. Pearson distance (Eq. (29))—$w = 0.1$ (left) and $w = 0.3$ (right).

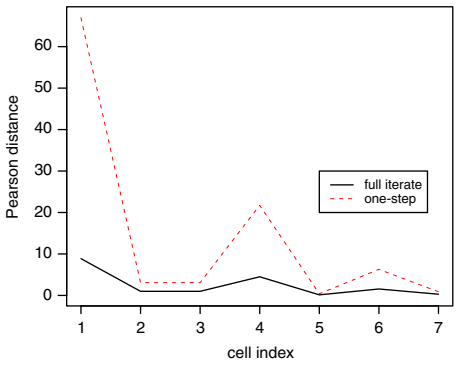

Fig. 7. Pearson distance (Eq. (29))—$w = 0.5$.

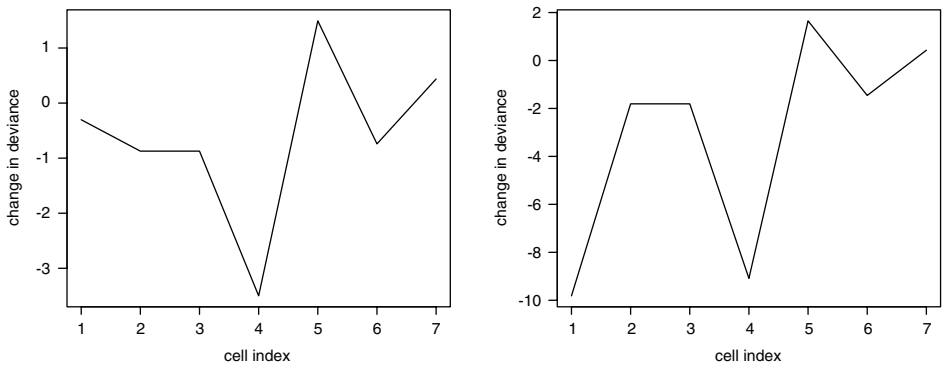

Fig. 8. Change in deviance (Eq. (32))—$w = 0.1$ (left) and $w = 0.3$ (right).

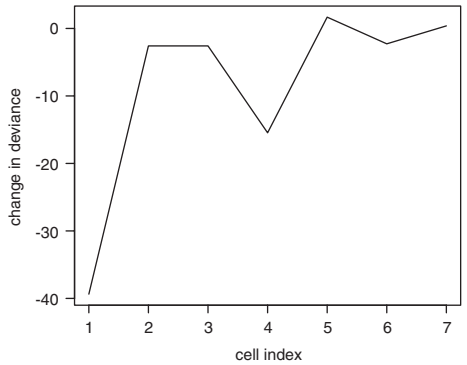

Fig. 9. Change in deviance (Eq. (32))—$w = 0.5$.

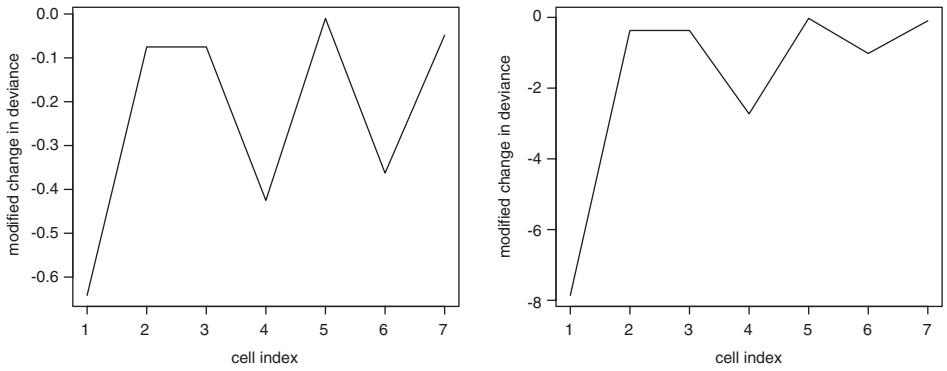

Fig. 10. Modified change in deviance (Eq. (33))—$w = 0.1$ (left) and $w = 0.3$ (right).

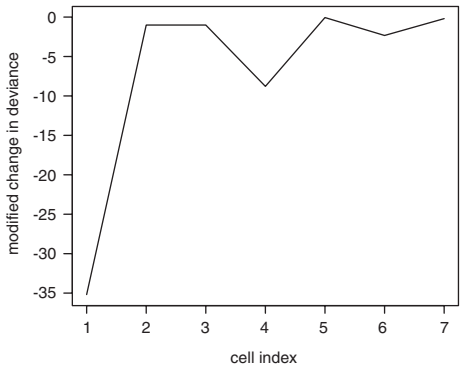

Fig. 11. Modified change in deviance (Eq. (33))—$w = 0.5$.

Table 1
Observations, fitted values and adjusted residuals

Department	Admission/male			Admission/female	
	Code	Yes	No	Yes	No
A	cell	1	7	13	19
	obs	512(−4.15)	313 (4.15)	89 (4.15)	19 (−4.15)
B	cell	2	8	14	20
	obs	353(−0.50)	207 (0.50)	17 (0.50)	8 (−0.50)
C	cell	3	9	15	21
	obs	120 (0.87)	205 (−0.87)	202 (−0.87)	391 (0.87)
D	cell	4	10	16	22
	obs	138 (−0.55)	279 (0.55)	131 (0.55)	244 (−0.55)
E	cell	5	11	17	23
	obs	53 (1.00)	138(−1.00)	94 (−1.00)	299 (1.00)
F	cell	6	12	18	24
	obs	22 (−0.62)	351 (0.62)	24 (0.62)	317 (−0.62)

Codes: obs = observations, cell = cell index.

a relatively large influence on the analysis. This in turn may indicate cells which give rise to unreliable prediction and estimation.

7.2. Log-linear model for independence

Table 1 displays the effect of gender on whether admitted into graduate school at the University of California at Berkeley for the fall 1973 session (see Agresti, 1990, p. 226) together with the corresponding adjusted residuals based on the log-linear model of conditional independence of whether admitted and gender, given department. Applicants were classified by A = admission status, G = gender, D = department, for the six largest graduate departments at Berkeley. The model of interest may be stated as

$$\log m_j = \mu + A + G + D + A : D + G : D,$$

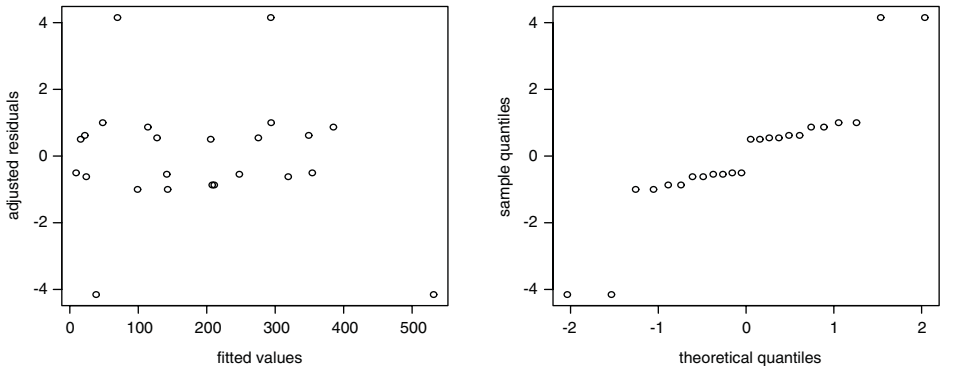

Fig. 12. Pearson residuals vs. fitted values (left) and Q–Q for adjusted residuals (right).

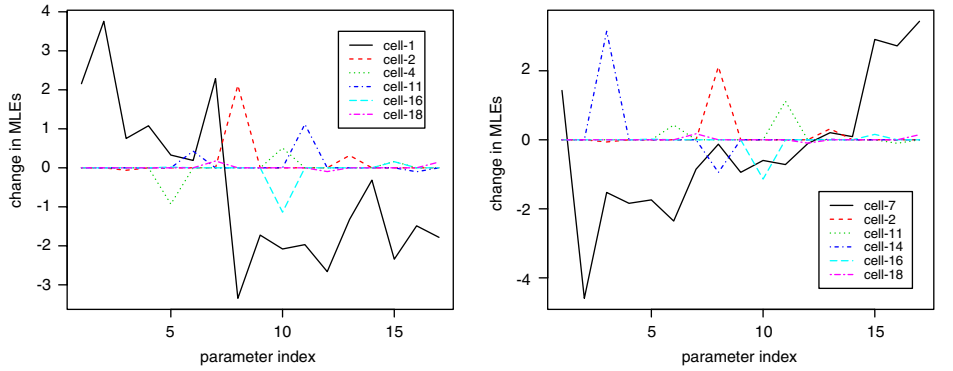

Fig. 13. Changes in MLEs.

where $m_j = n\pi_j$, the expected value of category j membership. We fitted this model to the data in Table 1 using R. The Pearson χ^2 statistic and the deviance were respectively found to be 19.9 and 20.2 based on 6 degrees of freedom, indicating a poor fit (Agresti, 1990). An analysis of residuals explains this. The plot of adjusted residuals against the fitted values and the normal Q–Q plot of the adjusted residuals (Fig. 12), reveal patterns in the residuals and show four very large residuals corresponding to observations from department A. This is compounded by the low chance (0.02) of a female being admitted to department A compared to that of a male (0.12). The variance for the Pearson residual for male admission to the department A is the lowest (0.04) while that for female admission in the same department is high (0.3).

Good diagnostic tools must identify the problematic observations corresponding to department A as influential. We introduced simple perturbations as defined by (8) to various cell probabilities of the type described by (35) and monitored the response of the ML estimates to these perturbations. Figs. 13 and 14 display changes in the parameter estimates when a selection of cells are perturbed. In these figures, cell 1, for example stands for the changes in ML estimates that occur when cell 1 is perturbed. Cell 1 contains the information

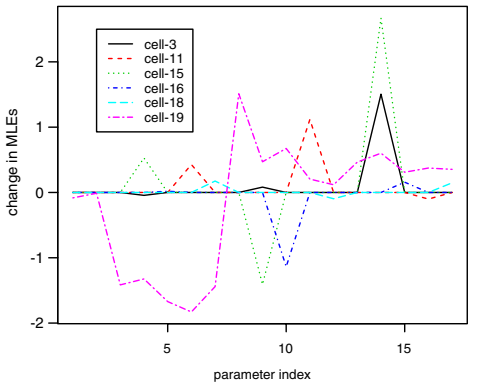

Fig. 14. Changes in MLEs.

about males that were admitted to department A while cell 7 gives the corresponding information on males whose admissions were turned down in the same department. It is clear that a perturbation of probabilities in the cells 1 and 7 result in the most deviant values of $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(w)$, which imply that these cells are influential. Other diagnostics derived in this paper produced quite similar conclusions. It can be seen (Fig. 14) that perturbing Cell 19 also causes substantial change in parameter estimates. It can also be observed from Fig. 14 that perturbing certain cells corresponding to departments other than A resulted in noticeable changes in certain parameter estimates, for example, perturbing cell 15 resulted in a large change in parameter corresponding to the interaction term between whether admitted (Y) and department (C) (given as "admY:depC" on the R output). We then see that perturbation analysis gives more insight into the data than does the residuals analysis only as done by Agresti (1990). One way of dealing with influential observations is to omit them from the analysis (Agresti, 1990). For example, omitting influential cases from the log-linear analysis results in an excellent fit with both the residual deviance and the Pearson statistic reducing to 2.7 at 5 df. However, removing these observations may conceal the information that they contain. Why is department A so different? Diagnostic analysis suggests an investigation of the department A to discover why it stands out on its own. Other departments like C whose corresponding cells resulted in noticeable changes on specific parameters when perturbed may be investigated as well.

8. Discussion

We have presented a new approach, namely perturbation, for identifying influential multinomial observations. The method has been used to derive several new diagnostic tools for identifying influential observations for the data that are dependant. Diagnostic tools that have been developed for many models are based on the one-step parameter estimates. These estimates are popularly used because they are obtained by a single step of Newton's method using ML estimates for the unperturbed model as the starting value and are hence less expensive to compute unlike their full iterative counterparts. In this paper it has been demonstrated

that one-step estimates are biased and diagnostic tools constructed using them tend to exaggerate the influence of individual cases, especially when the perturbed cases are influential. However, a comparison of diagnostic tools based on one-step estimates with those based on their full iterative counterparts, indicate that both these estimates lead to diagnostic tools that effectively identify the influential cases. Moreover, the one-step estimates led to the definition of new influence diagnostic tools for multinomial models that include influence curves, the average influence function and the Cook's curvature, both of which were shown to effectively identify the influential cases.

If it is desired to choose diagnostic tools for influential multinomial observations on the basis of our perturbation scheme, our computations indicate that a wide range of diagnostics presented here that include influence curves, the average influence function, the change in parameter estimates, likelihood distance, Cook's curvature and the modified change in deviance, all effectively identify influential cases. However, both the change in deviance and the Pearson distance were unable to detect the influential observation at small values of w, nevertheless, both these diagnostics correctly identified influential cases at reasonably high values of w. Undoubtedly, the new diagnostic tools extend and complement those based on case deletion proposed by Andersen (1992) and Nyangoma (2000).

As pointed out by a referee, it is possible to define perturbations other than the one studied in this paper. The most immediate example that comes to mind is to define w_j in Eq. (4) by

$$w_j = \begin{cases} w & \text{if } j \neq i, \\ w + \lambda/\pi_i & \text{if } j = i, \end{cases} \tag{37}$$

where $0 \leqslant w < 1$, $i, j = 1, \ldots, m$. Like the scheme studied in this paper, this perturbation affects the estimation by inducing a new perturbation, $\varphi_i(w) = w\pi_i / (w\pi_i + \lambda)$, that affects only the ith cell. For a fixed π_i, $\varphi_i(w)$ is a monotonic increasing function of w. This means that the results in this paper may be derived by considering limits as $w \to 0$ (instead of $w \to 1$ for the perturbation discussed in this manuscript). Under this scheme $w = 1$ represents the null perturbation. For fixed π but varying w, $\phi_i(w)$ forms a curve that is a reflection of the corresponding $\varphi_i(w)$ curve, about the line $w = 0.5$.

It has been established that the perturbation induced by conditioning on certain cells (Andersen, 1992) and those presented in this manuscript, all involve the simultaneous perturbation of the model assumptions about the probability of belonging to a cell by assigning them weights in a manner that leads to perturbation of only the ith cell. This manuscript thus unifies the perturbation theory for multinomial models that can be used to construct diagnostics for influential cell observations.

Acknowledgements

The first author gratefully acknowledges the invaluable discussions he had with Professor George Seber of the University of Auckland on the foundation ideas in this paper. The referees's comments gave new insights into the ideas presented in some sections of this manuscript.

References

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.
Andersen, E.B., 1992. Diagnostics in categorical data analysis. J. Roy. Statist. Soc. B 54, 781–791.
Andrews, D.F., Pregibon, D., 1978. Finding the outliers that matter. J. Roy. Statist. Soc. B 40, 85–94.
Basu, A., Basu, S., 1998. Penalised minimum disparity methods for multinomial models. Statist. Sinica 8, 841–860.
Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.
Birch, M.W., 1964. A new proof of the Pearson–Fisher theorem. Ann. Math. Statist. 35, 818–824.
Cain, K.C., Lange, N.T., 1984. Approximate case influence for proportional hazards regression model with censored data. Biometrics 40, 493–499.
Cook, R.D., 1977. Detection of influential observations in linear regression. Technometrics 19, 15–18.
Cook, R.D., 1979. Influential observations in linear regression. J. Amer. Statist. Assoc. 74, 169–174.
Cook, R.D., 1986. Assessment of local influence (with discussion). J. Roy. Statist. Soc. B 48, 133–169.
Cook, R.D., Weisberg, S., 1979. Finding influential observations in linear regression: a review. Technical Report 338, University Minnesota.
Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman & Hall, New York, London.
Elandt-Johnson, R.C., 1971. Probability Models and Statistical Methods in Genetics. Wiley, New York.
Escober, L.A., Meeker Jr., W.Q., 1992. Assessing influence in regression with censored data. Biometrics 48, 507–528.
Fung, W.-K., 1992. Some diagnostic measures in discriminant analysis. Statist. Probab. Lett. 13, 279–285.
Fung, W.-K., 1995. Diagnostics in linear discriminant analysis. J. Amer. Statist. Assoc. 90, 952–956.
Fung, W.-K., Kwan, C.W., 1997. A note on local influence based on the normal curvature. J. Roy. Statist. Soc. B 59, 839–843.
Haberman, S.J., 1973. The analysis of residuals in cross-classified tables. Biometrics 29, 205–220.
Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1985. Robust Statistics, the Approach Based on Influence Functions. Wiley, New York.
Hirschfeld, J., Heiken, A., 1963. Application of the Gc system in paternity cases. Amer. J. Human Genet. 15, 19–23.
Lawrance, A.J., 1991. In: Stahel, W., Weisberg, S. (Eds.), Directions in Robust Statistics and Diagnostics Part I. Springer, Berlin, pp. 140–157.
Lee, A.J., Nyangoma, S.O., Seber, G.A.F., 2002. Confidence regions for multinomial parameters. Comput. Statist. Data Anal. 39, 328–345.
Lee, S.-Y., Xu, L., 2004. Influence analyses of nonlinear mixed-effects models. Comput. Statist. Data Anal. 45, 329–342.
Nyangoma, S.O., 2000. Influential multinomial observations. Student 3 (3), 171–188.
Pregibon, D., 1981. Logistic regression diagnostics. Ann. Statist. 9, 705–724.
Rao, C.R., 1973. Linear Statistical Inference and its Applications. Wiley, New York.
Seber, G.A.F., Nyangoma, S.O., 2000. Residuals for multinomial models. Biometrika 87, 183–191.
Shi, L., 1997. Local influence in principal components analysis. Biometrika 84, 175–186.
Tang, N.-S., Wei, B.-C., Wang, X.-R., 2002. Detection of influential observations in nonlinear reproductive dispersion models. Aust. N. Z. J. Statist. 44, 185–192.
Weisberg, S., 1985. Applied Linear Regression. Wiley, New York.
Williams, D.A., 1987. Generalized linear model diagnostics using the deviance and single case deletions. Appl. Statist. 3 (2), 181–191.